

Test-time scaling of Opus 5 on robot arms

We gave Claude Opus 5 control of a pair of YAM arms and evaluated it on an [Inspect Robots](#) task: “Place the left block on top of the right block”.

We ran 15 trials across three effort levels (low, medium, and high; five per level), with fast mode enabled, and compared them with five runs of GPT-5.6 Sol at high reasoning effort.

Scoring rubric. Each run scored out of 100 across six weighted milestones.

Milestone	pts
Identifies the left block	10
Grasps the block	10
Lifts it over the other	20
Touches the other’s upper side	20
Success: stacked and standing	30
Accurate final report	10

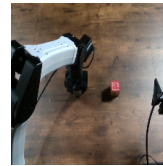
Milestones. Quote from Opus mid-run

identifies



“Overhead view shows two small blocks near the middle of the table: a blue-ish one left of centre and a red one right of centre. I’m swinging the left arm out over the left block at roughly 0.42 m radius, gripper pointing down at ~15 cm height, to get a wrist-camera look.”

grasps



“Block now sits centred in the jaw gap in the wrist view. Closing the gripper fully to test the grasp and read the residual thickness.”

lifts



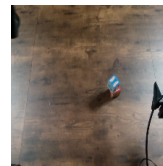
“Move was delta-clamped and only partly executed; the red block sits just off the top-right corner of my held block. Re-issuing slightly beyond the target to converge on being directly over it.”

touches



“Only a sliver of the red block still peeks above my carried block, so alignment looks good. Continuing a small ~2 cm descent, watching for the effort rise that means the carried block has landed.”

success



“The arm still blocks the overhead camera. Swinging it back toward its own base to clear the view of the blocks.”

Robotics capabilities of Opus 5 scale with effort levels

Figure 1. API cost per run vs weighted score; bold indicates the condition mean. Opus 5 at fast-mode rates (\$10/\$50 per MTok); Sol at \$5/\$30.

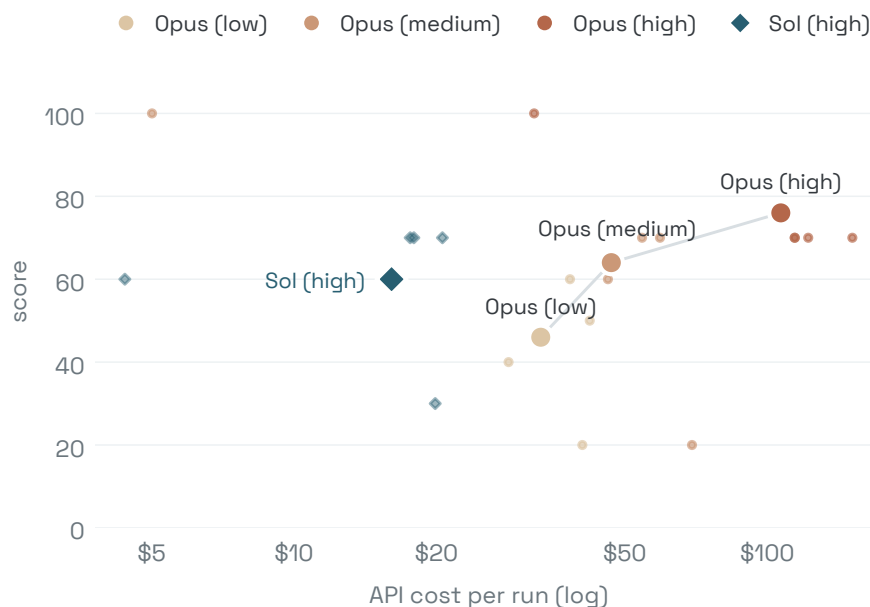
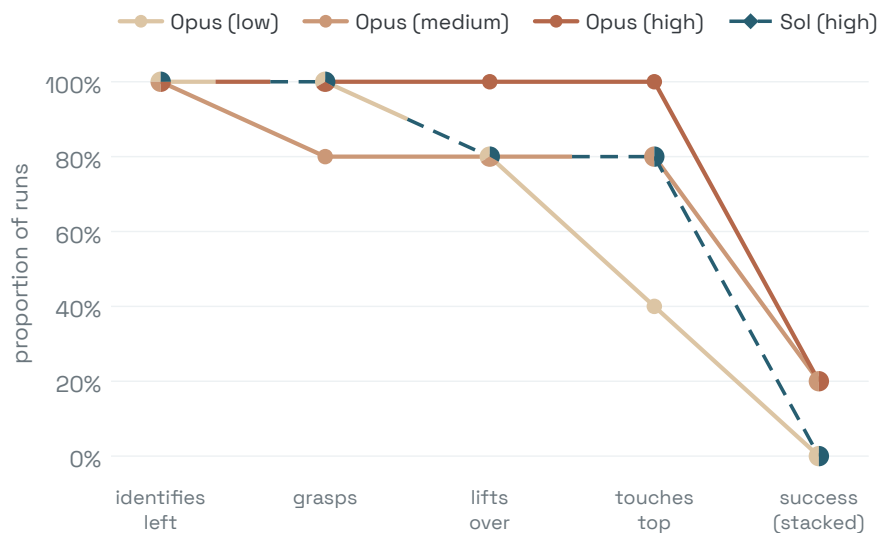


Figure 2. Milestone funnel: show how well each effort level performs on average at each rung of the milestone ladder. Where conditions tie exactly, the shared node and line show every tied color.



Technical Specifications

Arms	I2RT YAM bimanual arms (two 6-DoF arms with parallel grippers)
Cameras	fixed overhead Intel RealSense D435 + one Intel RealSense D405 on each wrist, 224×224 RGB, images attached to every observation (image horizon 2)
Models	claude-opus-5 (Anthropic Messages wire, fast mode) and gpt-5.6-sol (OpenAI Responses wire)
Effort	Opus swept low / medium / high via <code>output_config.effort</code> , with thinking on its adaptive default (effort steers its depth); Sol at high reasoning effort
Max turns	100 LLM calls per run
Max output tokens	16,000 per call
Actions	one move per LLM call (action horizon 1); arm speed capped at 10%
Harness	inspect-robots 0.53.0; collision guardrail off, joint-effort reporting on, rig-facts document in the system prompt; full wire capture

Additional plots

Figure A1. Mean weighted milestone score across the Opus 5 effort ladder (max 100), with the GPT-5.6 Sol reference at right.

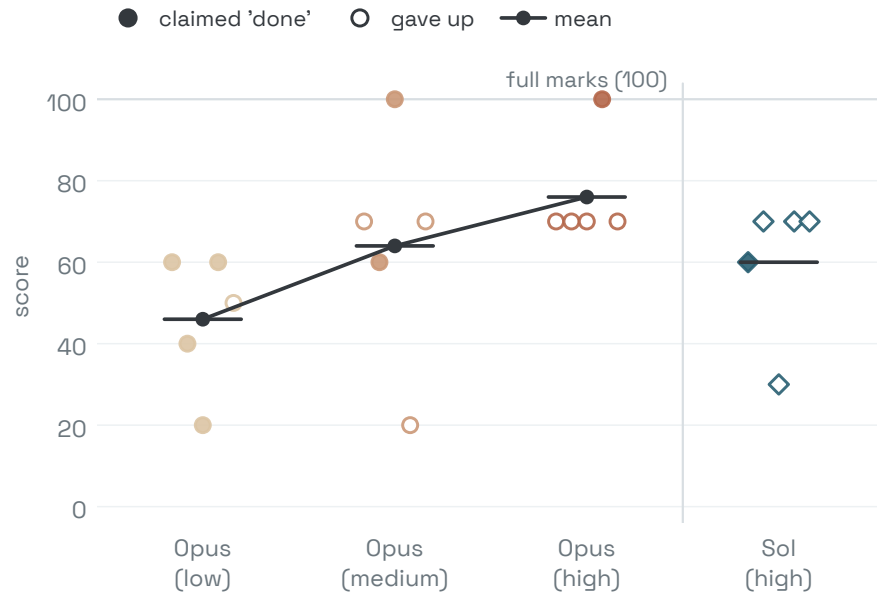


Figure A2. Output tokens vs weighted score. Bold indicates mean.

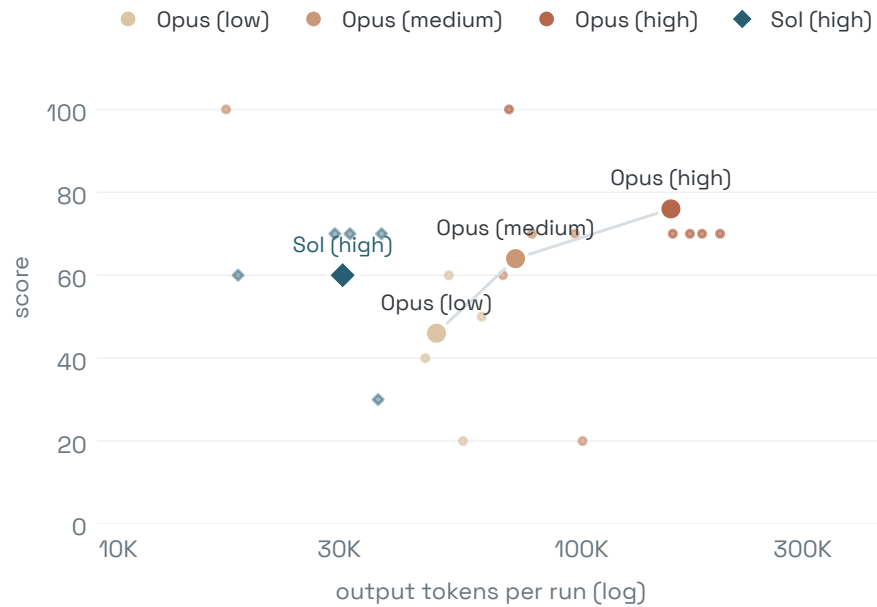


Figure A3. Reasoning tokens per run vs weighted score.

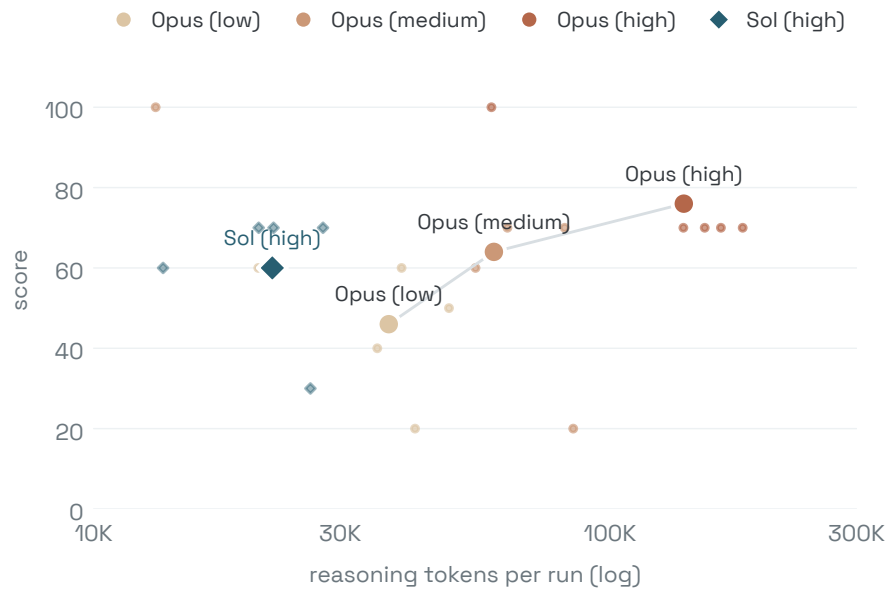


Figure A4. Score vs wall-clock duration.

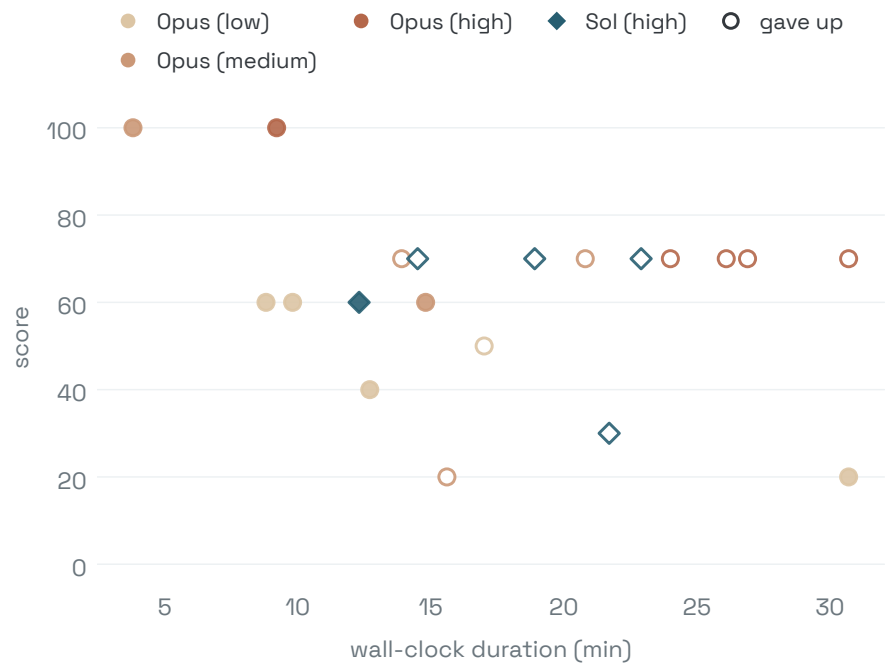
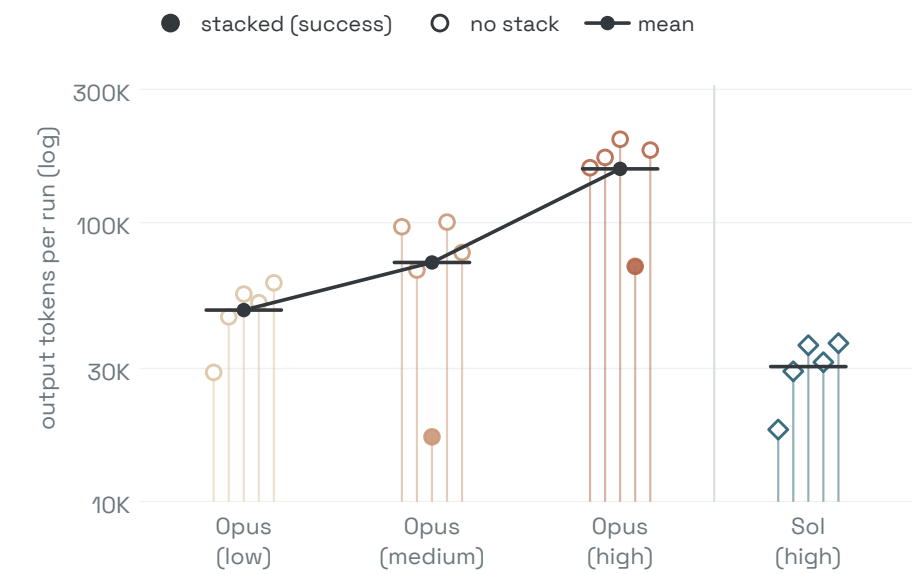


Figure A5. Total output tokens per run (generation including reasoning tokens). Excludes cache traffic.



All runs

Model	Effort	Score	Time (min)	Transcript	Video	Rerun
Opus 5	high	100	9.2	view transcript	watch	download (rerun)
Opus 5	high	70	24.0	view transcript	watch	download (rerun)
Opus 5	high	70	26.1	view transcript	watch	download (rerun)
Opus 5	high	70	26.9	view transcript	watch	download (rerun)
Opus 5	high	70	30.7	view transcript	watch	download (rerun)
Opus 5	medium	100	3.8	view transcript	watch	download (rerun)
Opus 5	medium	70	13.9	view transcript	watch	download (rerun)
Opus 5	medium	70	20.8	view transcript	watch	download (rerun)
Opus 5	medium	60	14.8	view transcript	watch	download (rerun)
Opus 5	medium	20	15.6	view transcript	watch	download (rerun)
Opus 5	low	60	8.8	view transcript	watch	download (rerun)
Opus 5	low	60	9.8	view transcript	watch	download (rerun)
Opus 5	low	50	17.0	view transcript	watch	download (rerun)
Opus 5	low	40	12.7	view transcript	watch	download (rerun)
Opus 5	low	20	30.7	view transcript	watch	download (rerun)
GPT-5.6 Sol	high	70	14.5	view transcript	watch	download (rerun)
GPT-5.6 Sol	high	70	18.9	view transcript	watch	download (rerun)

Model	Effort	Score	Time (min)	Transcript	Video	Rerun
GPT-5.6 Sol	high	70	22.9	view transcript	watch	download (rerun)
GPT-5.6 Sol	high	60	12.3	view transcript	watch	download (rerun)
GPT-5.6 Sol	high	30	21.7	view transcript	watch	download (rerun)

Opus can handle swapped arms

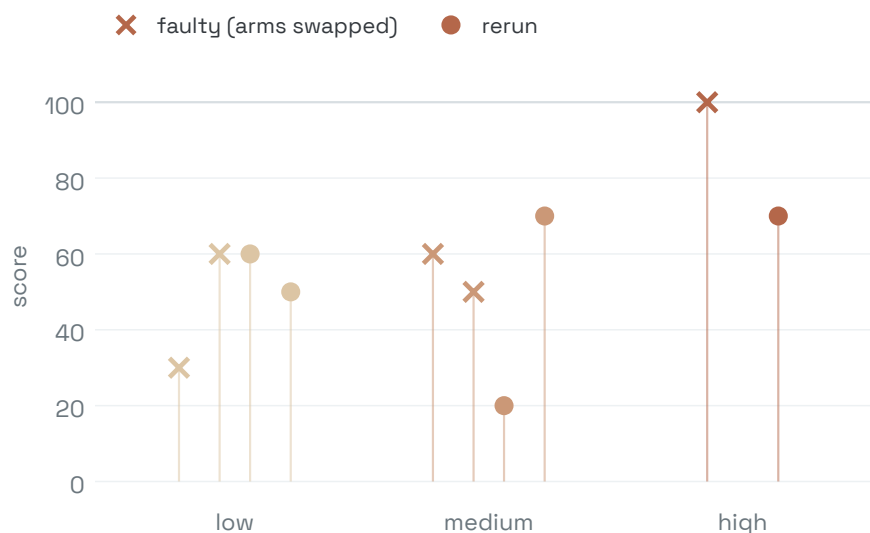
In some of our early runs the two arms were swapped (the harness had stored the left arm as “right” and the right arm as “left”). We reran those trials, but it turned out that Opus could handle the swap and figure the problem out itself.

“Interesting: moving the left arm changed the image labelled ‘right_cam’, so the wrist camera labels appear swapped.” — run 09a67d86 (medium)

“Camera labels appear swapped: ‘right_cam’ is actually the left arm’s wrist view, showing the red block at the left edge.” — run 078578dd (low)

“The arm that moved appears on the RIGHT side of the overhead image, so the top view is mirrored relative to the arm frame; I’m swinging the left arm’s base yaw by -0.6 rad to confirm the lateral direction...” — run ade192ba (high)

Figure A6. Faulty omen-2 runs (arms swapped) vs their reruns. Opus only.



Model	Effort	Score	Time (min)	Transcript	Video	Rerun
Opus 5	high	100	9.6	view transcript	watch	download (rerun)
Opus 5	medium	60	11.5	view transcript	watch	download (rerun)

Model	Effort	Score	Time (min)	Transcript	Video	Rerun
Opus 5	medium	50	16.5	view transcript	watch	download (rerun)
Opus 5	low	60	11.5	view transcript	watch	download (rerun)
Opus 5	low	30	19.1	view transcript	watch	download (rerun)

Interesting failure mode

Often, the models claimed to be done when they had built a tower lying on the table, as in the frame below.

